

FIG. 1

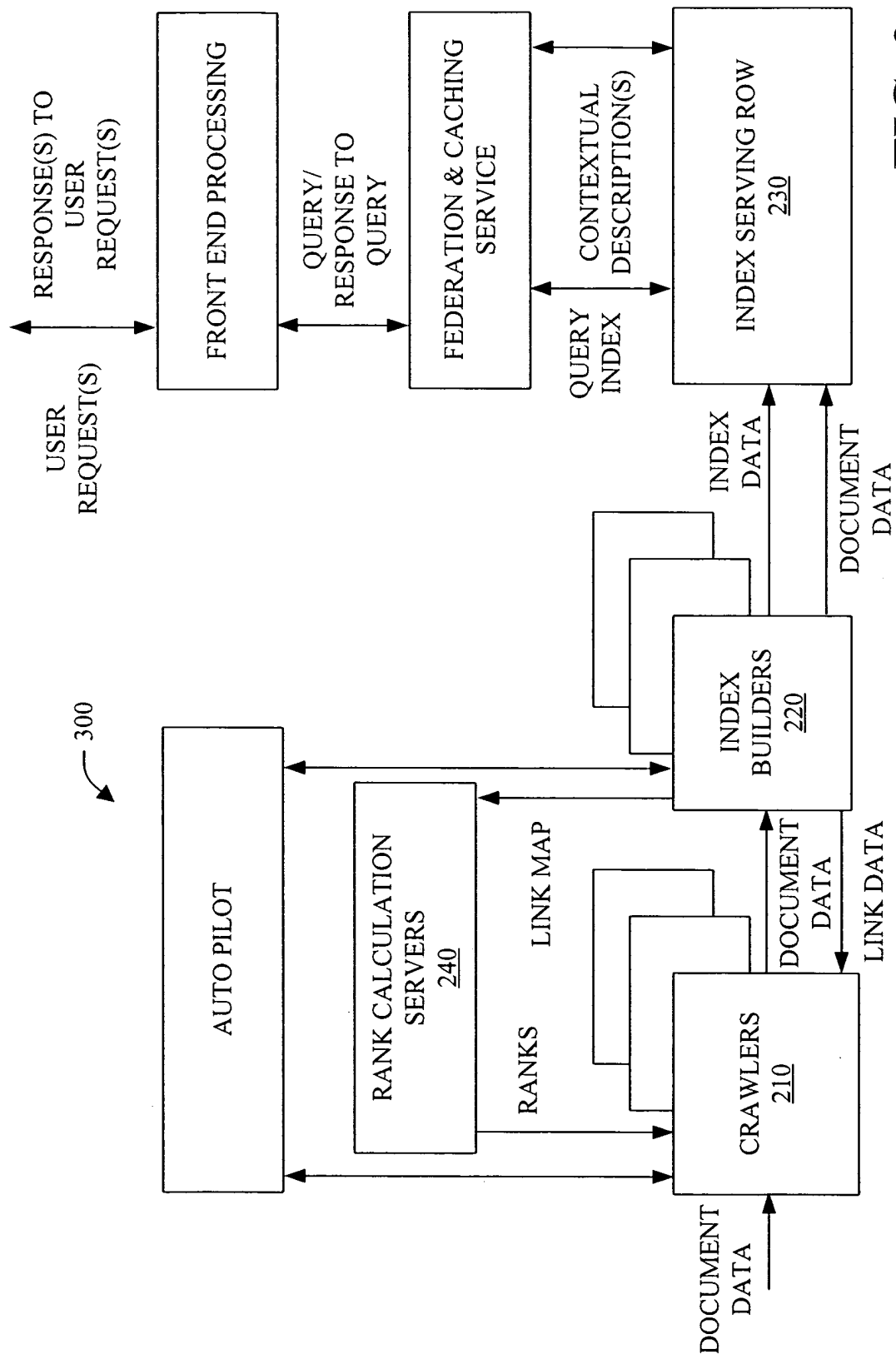


FIG. 2

HEADER:

4 BYTE MAGIC NUMBER
4 BYTE FILE HEADER SIZE
4 BYTE FILE VERSION SIZE
4 BYTE DOCUMENT HEADER SIZE
4 BYTE COUNT OF INDEX ELEMENTS
4 BYTE INDEX ELEMENT SIZE
ARRAY OF INDEX OF OFFSETS

HEADER**DOCUMENT:**

16 BYTE IP ADDRESS FROM WHERE DOCUMENT WAS DOWNLOADED
(IPV6 SUPPORT)

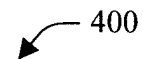
8 BYTE TIME DOWNLOADED (FILETIME)
8 BYTE DURATION OF DOWNLOAD (FILETIME)
4 BYTE CRC-32 OF UNCOMPRESSED DOC
4 BYTE COUNT OF BYTES IN THE UNCOMPRESSED DOC
4 BYTE COUNT OF BYTES IN THE COMPRESSED DOC
2 BYTE COUNT OF BYTES IN THE URL
1 BYTE LINK DEPTH
1 BYTE FLAGS (USED TO DENOTE IF THE DOC IS COMPRESSED)
4 BYTE COUNT OF BYTES FOR ADDITIONAL METADATA
URL

ADDITIONAL METADATA

RAW DOCUMENT WITH HTTP HEADERS MODIFIED ONLY BY THE
REMOVAL OF "TRANSFER CODING: CHUNKED"
DOCUMENT

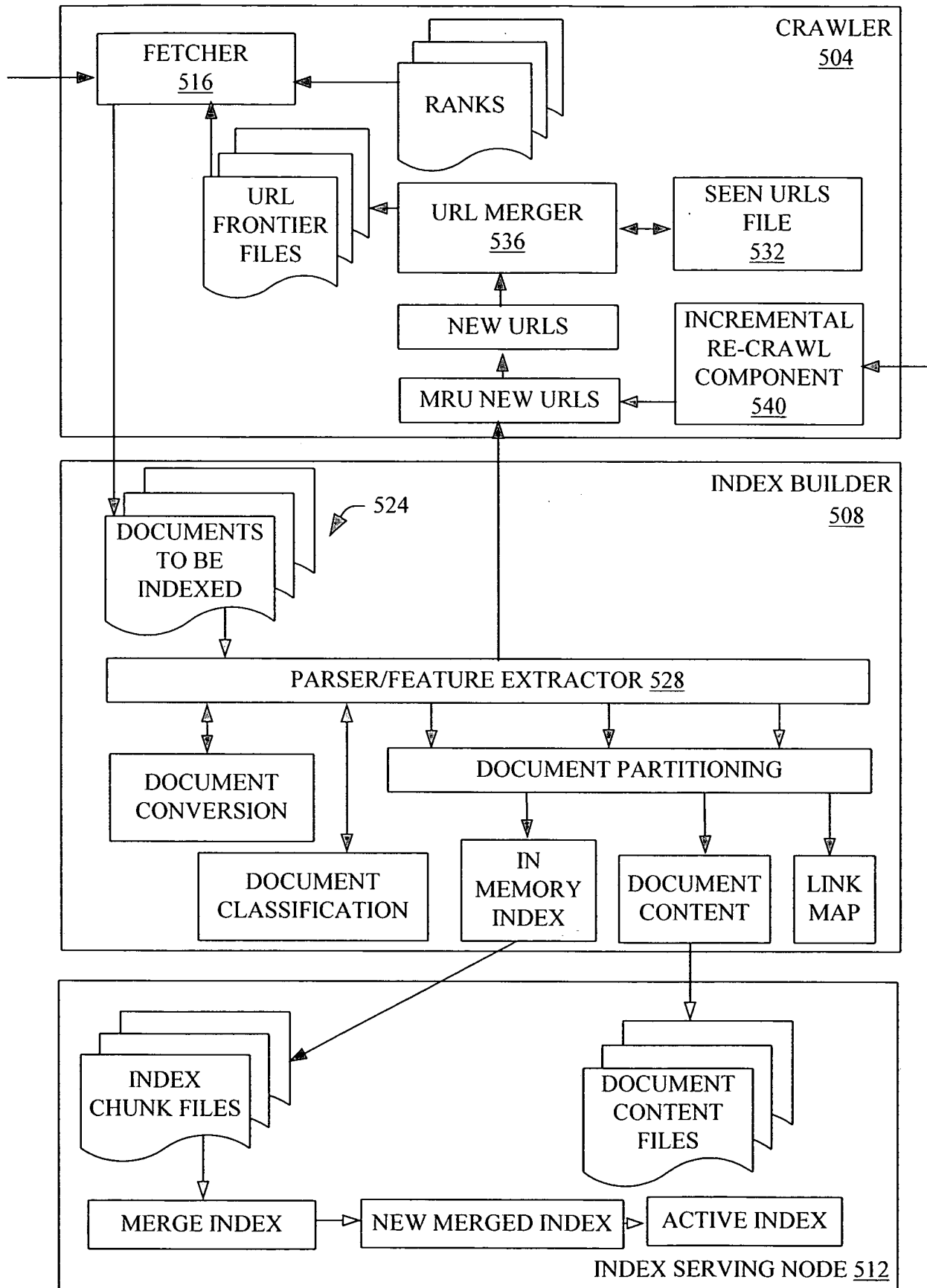
FIG. 3

400



TYPE ID
URL
HASH OF URL
CHUNK ID (CHUNKS ARE EXPLAINED LATER)
RANK (STATIC RANK FOR A PAGE)
LINK DEPTH
FINGERPRINT (KEY USED TO TELL IF PAGE HAS CHANGED)
HASH OF SOURCE URL OF ANCHOR TEXT
ANCHOR TEXT STRING

FIG. 4



600

HEADER <u>604</u>
OFFSET(S) <u>608</u>
DOCUMENT FILE ₁ <u>612₁</u>
• • •
DOCUMENT FILE _p <u>612_p</u>

FIG. 6

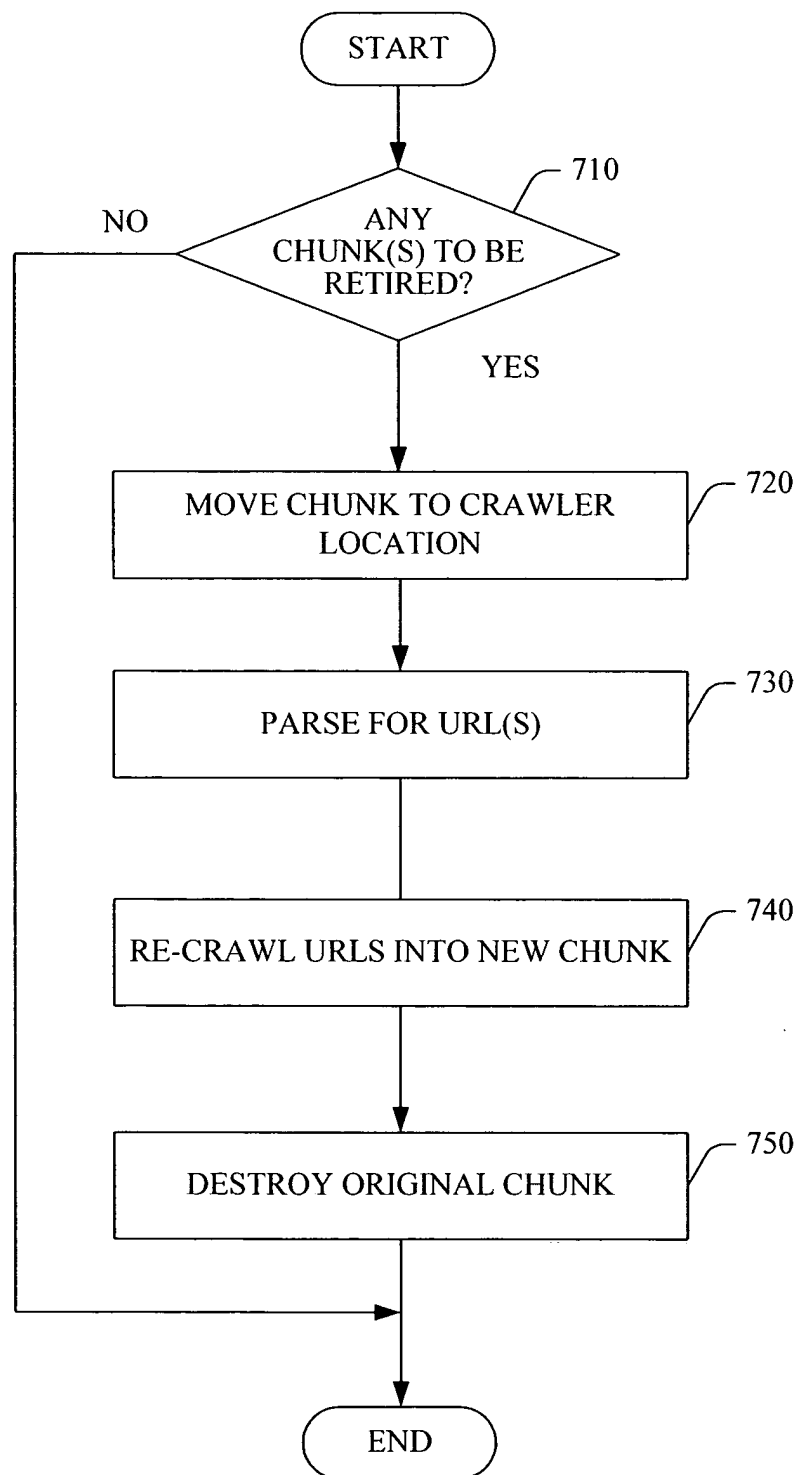


FIG. 7

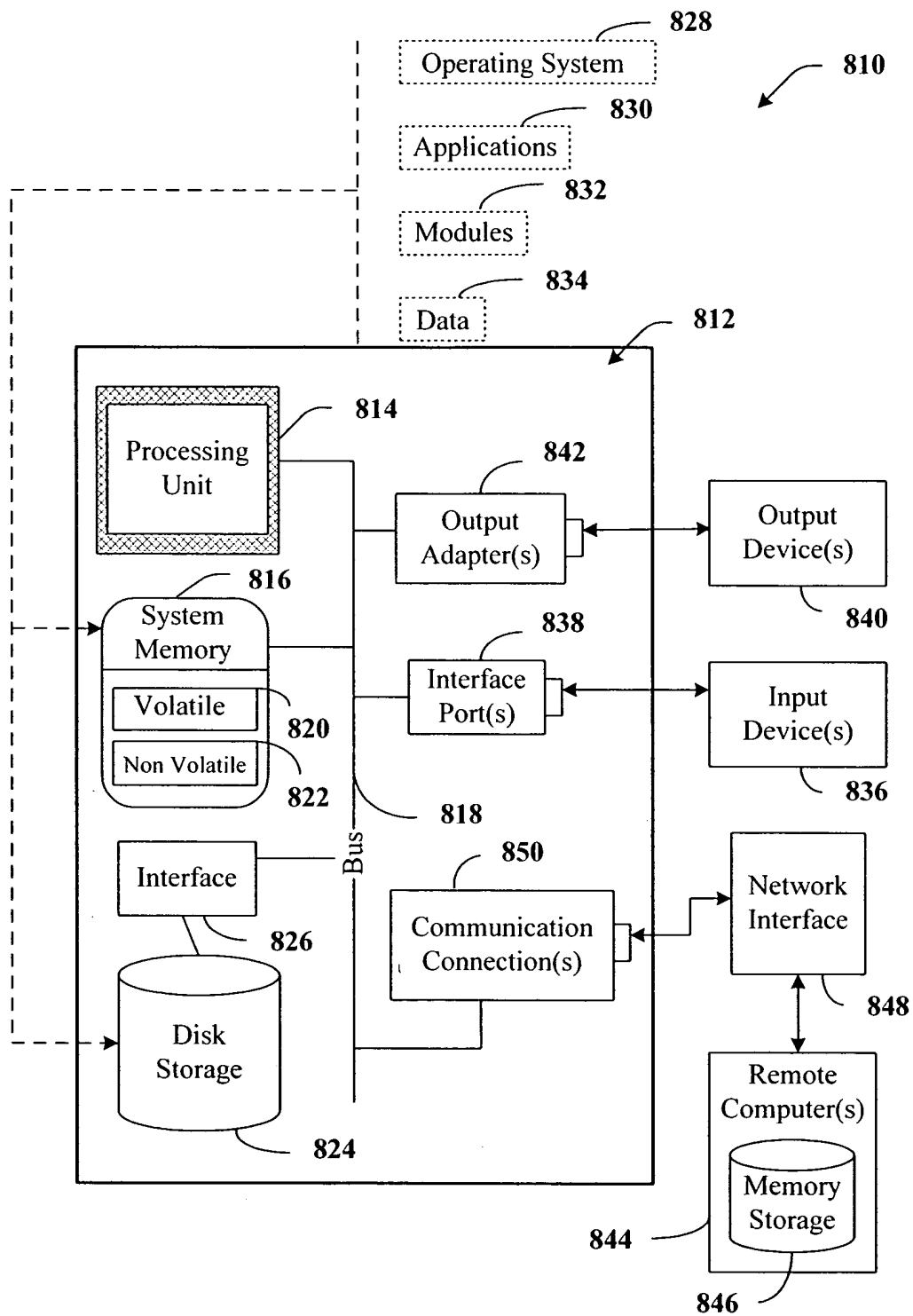


FIG. 8